ФАКУЛЬТЕТ ЛИНГВИСТИКИ

АННОТИРОВАНИЕ КОРПУСОВ ТЕКСТОВ

О.И. Бабина

В прикладной лингвистике термин «аннотирование» текстов трактуется как сообщение определенной дополнительной лингвистической информации о тексте, что реализуется посредством его разметки в соответствии с определенной концепцией или теорией. Концепция, в рамках которой производится аннотирование, может определить задачи автоматической обработки текста и устанавливать новые стандарты. Аннотирование корпуса текстов является предварительным условием для многих методов машинного обучения при решении задач автоматической обработки текста.

Разметка может осуществляться на различных уровнях:

1) *лексико-семантическом*: слова текста, как правило, полнозначные, соотносятся с заданной ограниченной базой дескрипторов, которые репрезентируют определенные смыслы, описанные в тезаурусе или онтологии. К семантическому виду разметки также относят определение предикатно-аргументной структуры текста;

2) *грамматическом* (в некоторых аннотационных схемах отдельно выделяются уровни морфологической и синтаксической разметки): предложения текста представляются в виде деревьев, в узлах которых располагаются слова с приписанными им метками об их морфологической форме, либо дерево может представлять собой абстрактную конструкцию, в которой лишь узлы-«листья» заполнены словами, остальные же узлы предоставляют информацию о свойствах соответствующих «листьев». Обобщенная структура такого дерева и предоставляемая в узлах или пометках дуг информация (о синтаксической функции, части речи узлов дерева и т.д.) определяется аннотационной схемой, лежащей в основе синтаксической разметки. Деревья могут отображать глубинно-синтаксический или поверхностно-синтаксический уровень предложения. Также грамматическая разметка предложения может включать аннотирование тема-рематических отношений;

3) *дискурсном*: чаще всего, сводится к выявлению и разметке анафорических и кореферентных отношений в тексте. При этом формирование анафорических связей рассматривается не только для традиционно выступающих в роли анафоров местоимений, но также другие лексические элементы, так называемые «анафорические дискурсивные связки» (*otherwise, instead, furthermore* и т.д.).

Ряд исследователей производят попытки создать универсальную схему аннотирования, которая бы покрывала все эти уровни.

На сегодняшний день не существуют автоматической системы аннотирования текстов, так как естественный язык представляет собой «гибкую» систему, с трудом поддающуюся формализованному описанию в рамках какой-либо из теорий. Аннотирование корпусов текстов осуществляется вручную либо автоматизирована. В последнем случае используются морфологические и синтаксические анализаторы, тезаурусы и онтологии. Однако традиционно окончательным арбитром в процессе аннотации текстов остается человек, и автоматизированная процедура аннотирования корпуса текстов с необходимостью включает в себя этап интер- или постредактирования.

Перед началом любого аннотирования необходимо определиться с общей концепцией аннотирования – определить, какая информация будет включена в метки, присваемые в ходе аннотирования. Предлагаемая нами автоматизированная процедура включает двухуровневую разметку текста: 1) на лексико-грамматическом уровне (для отдельных слов определяется принадлежность к части речи и парадигматическая форма); 2) на уровне предикатно-аргументной структуры (выделяются предикаты и помечаются с помощью инвентаря валентностей их роли). Для автоматизации этой процедуры используются приложения программного комплекса *LingAssistant*.

Предварительным этапом для аннотирования корпуса является построение лексикона, которое начинается с автоматического построения словника на основе корпуса текстов, сортированного по окончаниям.

Далее словник подвергается автоматическому анализу, и выявляются парадигматические кластеры словоизменительных квазисуффиксов, характерных для данного языка. Результатом данного этапа является список квазисуффиксов (разделенных точкой), предположительно составляющих одну парадигму, причем каждой такой парадигме приписывается список основ из словника, изменяющийся согласно данной парадигме. Например,

ing.ed.s: *draft, end, surround, …*

В случае если какая-то из основ не соответствует данной парадигме и оказалась в списке случайно, аннотатор может удалить ее из списка. Аналогично, может быть удалена вся парадигма, если она не существенна для определения закономерностей в языке.

Далее аннотатор вручную приписывает каждому из квазисуффиксов лексико-грамматическую метку (или список меток, в случае омонимичных парадигматических форм), которая автоматически присваивается всем соответствующим словоформам. Например, квазисуффиксу *-ing* присваивается метка ~Pger~Pg (где ~Pger – глагол (предикат), герундий, ~Pg – глагол, активное причастие). Словоформы основ из списка с присвоенными метками заносятся в лексикон, а выявленные квазисуффиксы используются на последующем этапе.

Далее производится автоматический отбор словоформ с квазисуффиксами первого порядка (условно соответствующих словообразовательным суффиксам, в отличие от предыдущего этапа, сконцентрированного на выявлении квазисуффиксов второго порядка – словоизменительных). Отбор осуществляется на основе баз знаний квазисуффиксов первого и второго порядка, а также правил их соединения [1].

Этап построения лексикона был осуществлен на материале английского языка, однако с небольшими затратами может быть реализован для других флективных языков или языков с достаточно сильным флективным компонентом.

Далее, на основе составленного лексикона, осуществляется собственно процедура автоматической аннотации текста. В основе аннотирования лежит процедура автоматического анализа [2, 3], дополненная эвристиками для обработки словоформ, не указанных в словаре:

1) *Токенизация*: выделение крупных блоков текстов на основе пунктуации и табуляции, а также списка стоп-слов (включающего служебные слова) – такие списки часто применяются в информационном поиске;

2) *Частичный лексико-грамматический анализ*: представляет собой разметку слов с использованием меток, хранящихся в лексиконе. Словоформам, не включенным в лексикон, присваиваются метки на основании эвристических правил, учитывающих контекст. В случае, когда словоформе приписывается список меток, посредством составленных на основе данных исследовательского корпуса продукционных правил производится разрешение неоднозначности. Результатом данного этапа является текст, каждому слову которого присвоена соответствующая лексико-грамматическая метка или метка ~UKN (форма неизвестна).

3) *Синтаксический анализ*: включает восходящую обработку данных, в ходе которой группы слов объединяются в блоки, идентифицируются их вершины и блоки классифицируются на основе лексико-грамматических меток вершин. Для выделения блоков используются метод распознавания образцов, дополненный правилами-ограничениями, специфичных для отдельных типов блоков. На выходе элементы текста объединены в группы, помеченные соответствующей синтаксической меткой.

4) *Семантический анализ*: состоит из идентификации предикатных единиц и определения семантических зависимостей – по правилам полученные на предыдущем этапе синтаксические блоки соотносятся с семантическими ролями предиката на основе своей синтаксической структуры (тип блока, предлог в инициальной позиции);

5) *Восстановление кореференции*: представляет собой эвристику для поиска существительных, вступающих в отношение кореференции (соответствующие существительные помечаются индексами), и анафоров для местоимений.

Результатом данной процедуры является набор предикатно-аргументных структур, имеющих следующий вид (для простоты лексико-грамматические метки опущены):

(Pg1 «culturing»
1 «the | host$_1$ | cell | of | claim | 6» // <dir-object>
2 «conditions | for | the | expression | of | the | protein$_2$» // <cond>[under])
(Pg2 «recovering»
1 «protein$_2$» // <dir-object>
2 «host | cell | culture$_1$» // <source>[from])

Результат анализа далее подвергается постредактированию.

Этап автоматического анализа более зависим от языка, чем построение лексикона, так как правила и эвристики обусловлены структурой определенного языка. Однако во многих европейских языках имеются схожие черты, поэтому правила, составленные для одного языка, могут быть отредактированы и вновь использованы для обработки новой предметной области или корпуса на другом языке.

Библиографический список

1. Бабина, О.И. Автоматическая первичная парадигматическая идентификация словоформ на основе анализа их морфологической квазиструктуры. / О.И. Бабина. // Актуальные проблемы теоретической и прикладной лингвистики: материалы Международной научной конференции (Челябинск, 10-13 декабря 2007 г.) / Отв. редактор О.А. Турбина. – Челябинск: Изд-во ЮУрГУ, 2007. – Ч. 1. – С. 42-45.

2. Sheremetyeva, S. Natural Language Analysis of Patent Claims. / S. Sheremetyeva. // Proceedings of the Workshop on Patent Corpus Processing. – Sapporo, Japan. – July 2003. – Pp. 66-73.

3. Бабина, О.И. Построение модели извлечения информации из технических текстов: дис. … канд.филол.наук / О.И. Бабина. – Челябинск, 2006. – 235 с.